**Exploring the Novamag database**

After importing all the .json files and removing any corrupt entries, we have a database of **1668** compounds. We have **6** failed imports due to problems with the original json files.

The material properties are split into subcategories: chemistry, crystal, thermodynamics, magnetism and additional information. For now I have focused on chemistry, crystal and magnetism properties.

Total number of imported features is: **32**

Total number of features with at least one NaN value is: 30. Therefore only 2 complete feautues (one of which is ‘chemical formula’).

We drop all features with more that 1650 NaN values i.e. less than 18 results. We will drop more for any machine learning we do. We find **14** features that have more than 1650 NaN, so this leaves us with **18** remaining features.

Of these features, **12** hold categorical data and **6** hold numerical data.

One of the categorical columns is ‘types of anisotropy’ which can take three values ‘easy axis’, ‘easy plane’ and ‘easy cone’. This low cardinality feature is easily encoded.

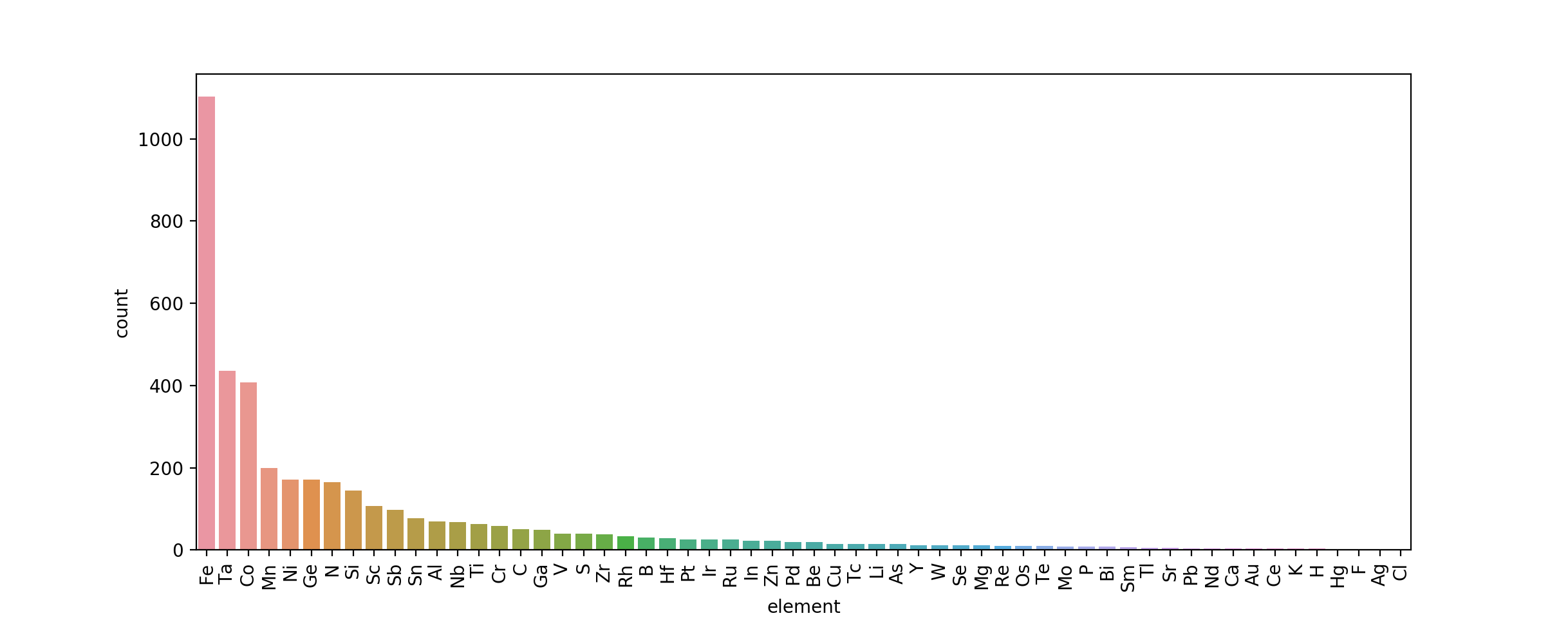
Of these 18 features, **13** are more than 60% populated with data. Saturation magnetization is available for more than 99% of the data and would make a good target for any machine learning.

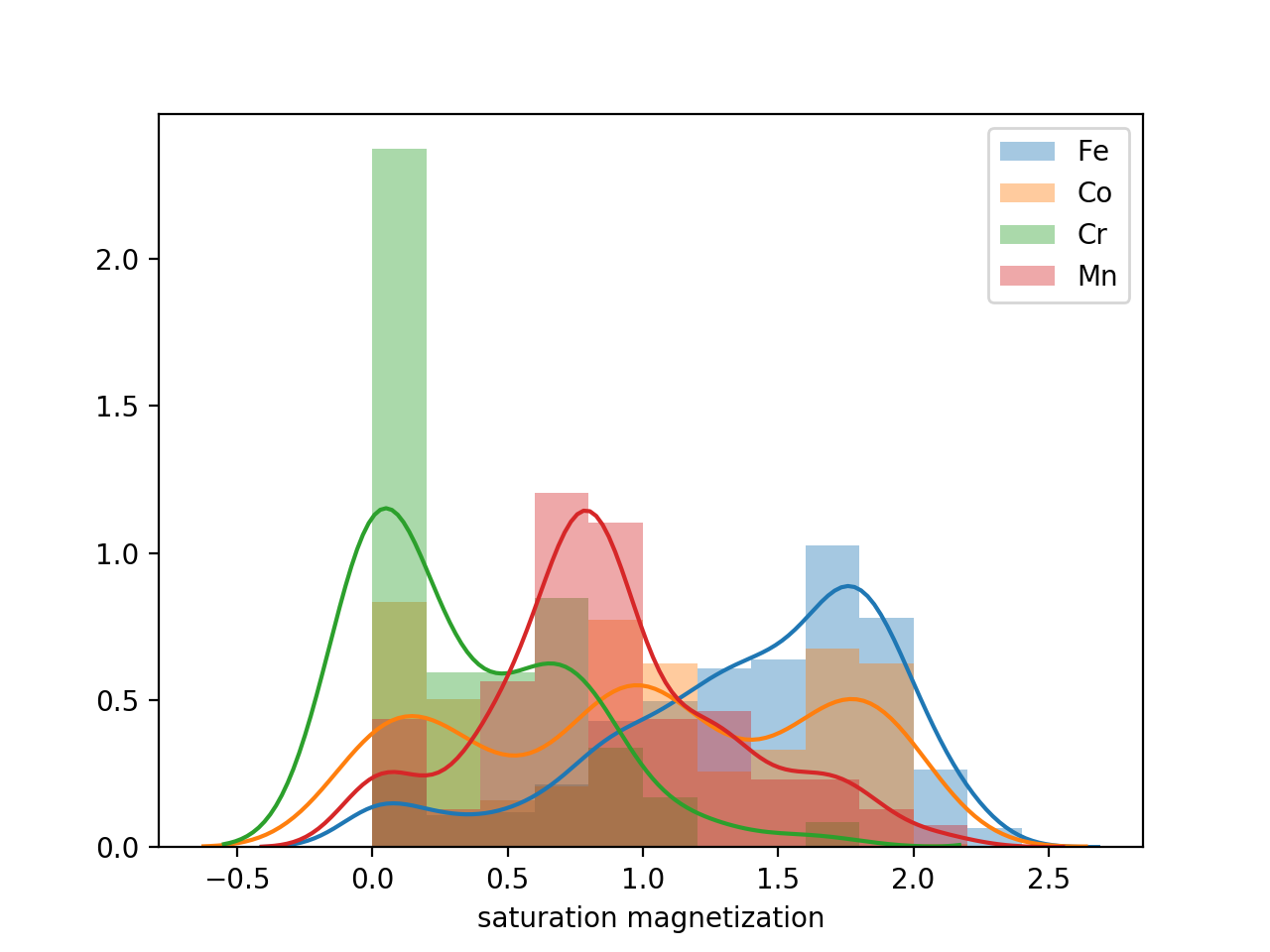
Unfortunately, of the 1668 compounds, **791** are unique compounds. Which is quite a major problem for any machine learning. Co2Fe12Ta2 is the most common duplicate, occurring 25 times. I haven’t removed these duplicate compounds yet, as I need to investigate the best way to merge their values into a single value.

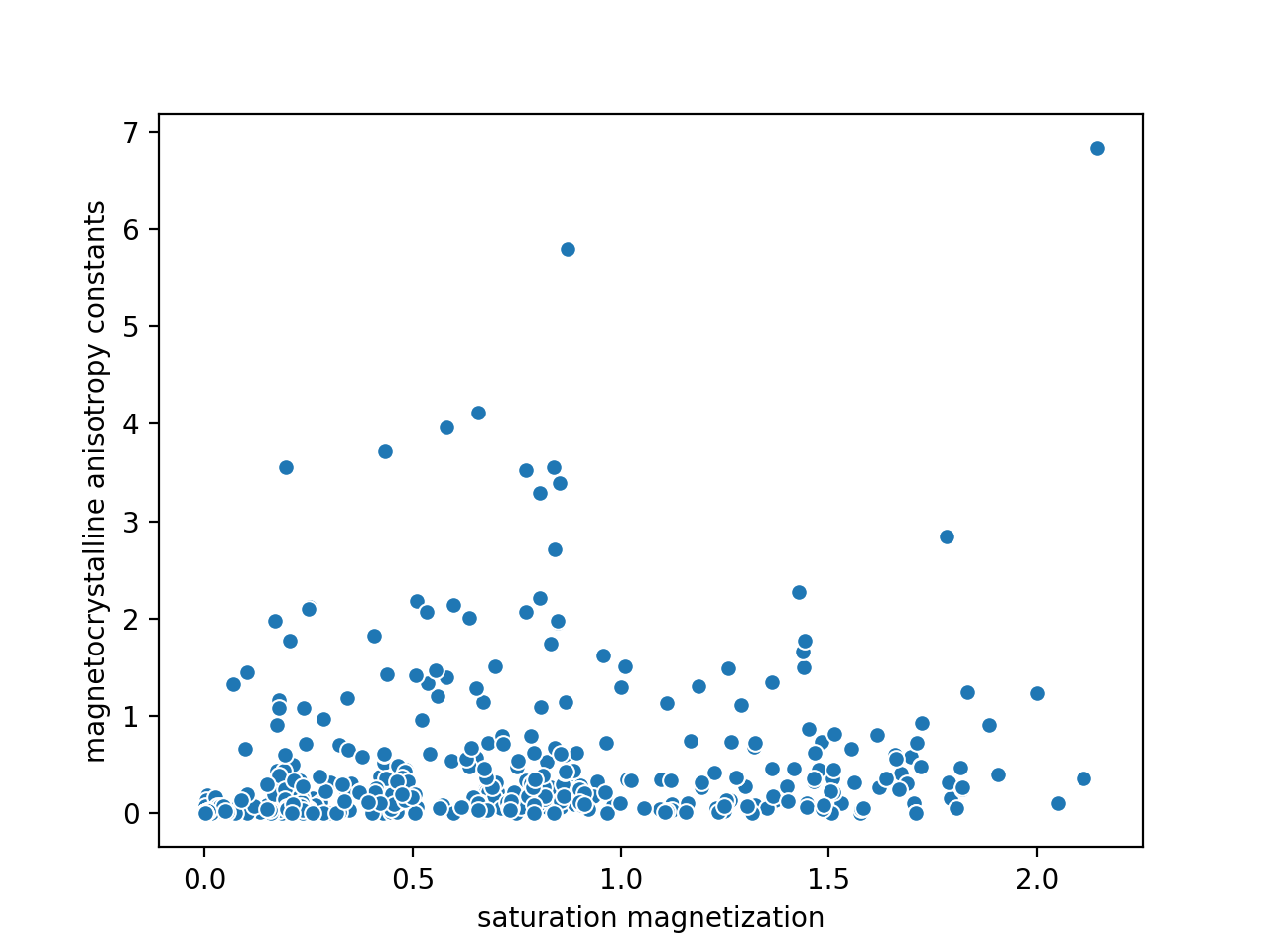
The number of binary compounds is **1004**

Ternary compunds is **647**

Quaternary compounds is **14**

The most common element in the compounds is **Fe with 1102 occurrences**.

We can plot the distribution of saturation magnetization for compounds grouped by the elements they contain. Perhaps unsurprisingly, compounds that contain Fe tend to have a higher saturation magnetisation and compounds that contain Cr have a lower saturation magnetisation. Note that the lines are Kernel density estimates.

We can compare two separate features in search of a correlation. The Spearman’s rank coefficient of abs(K) with Ms is 0.73.

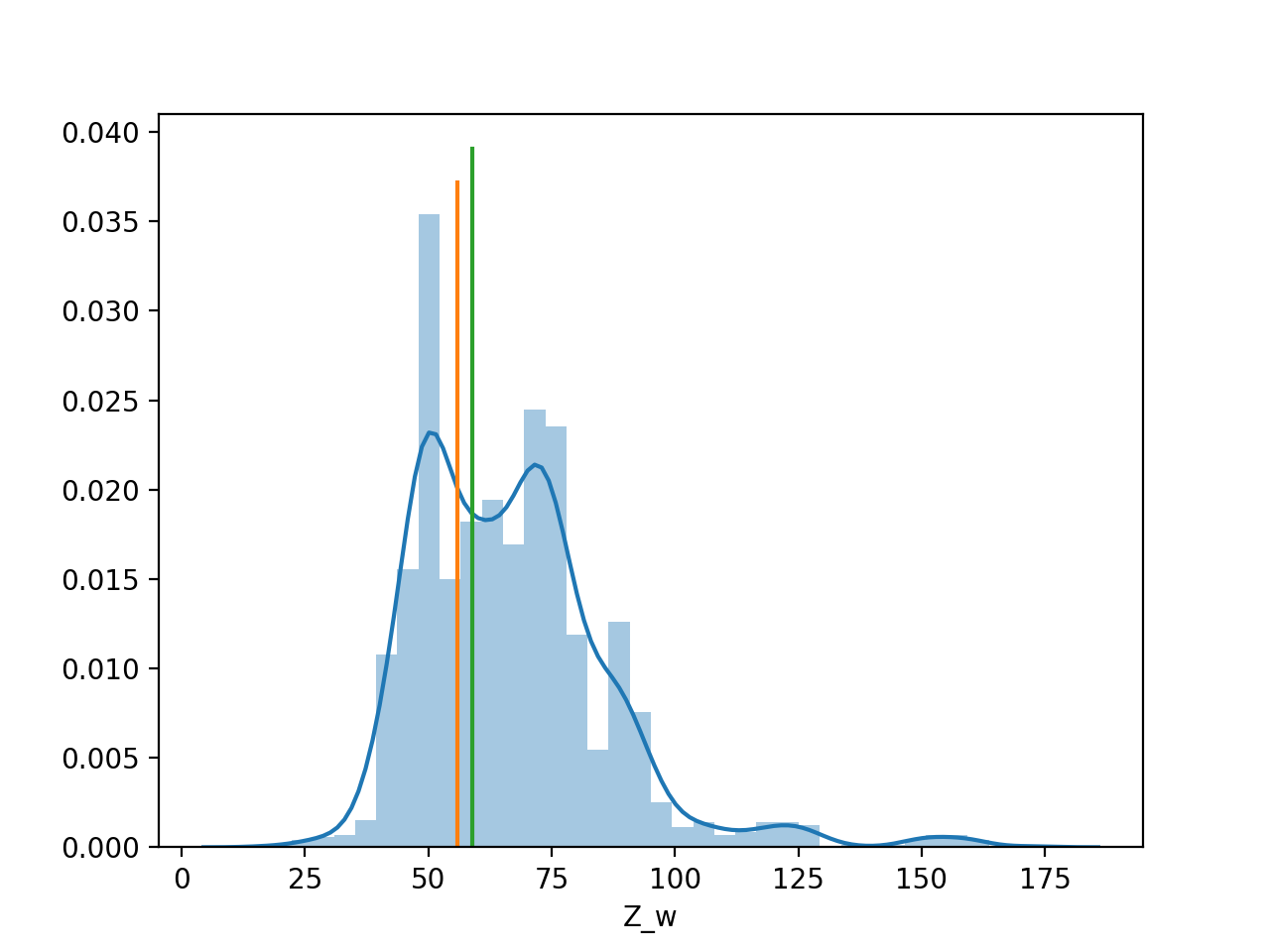
Using the chemical formula, we can design a number of features based on readily available data from the periodic table which composition weighted values of the comprising elements.

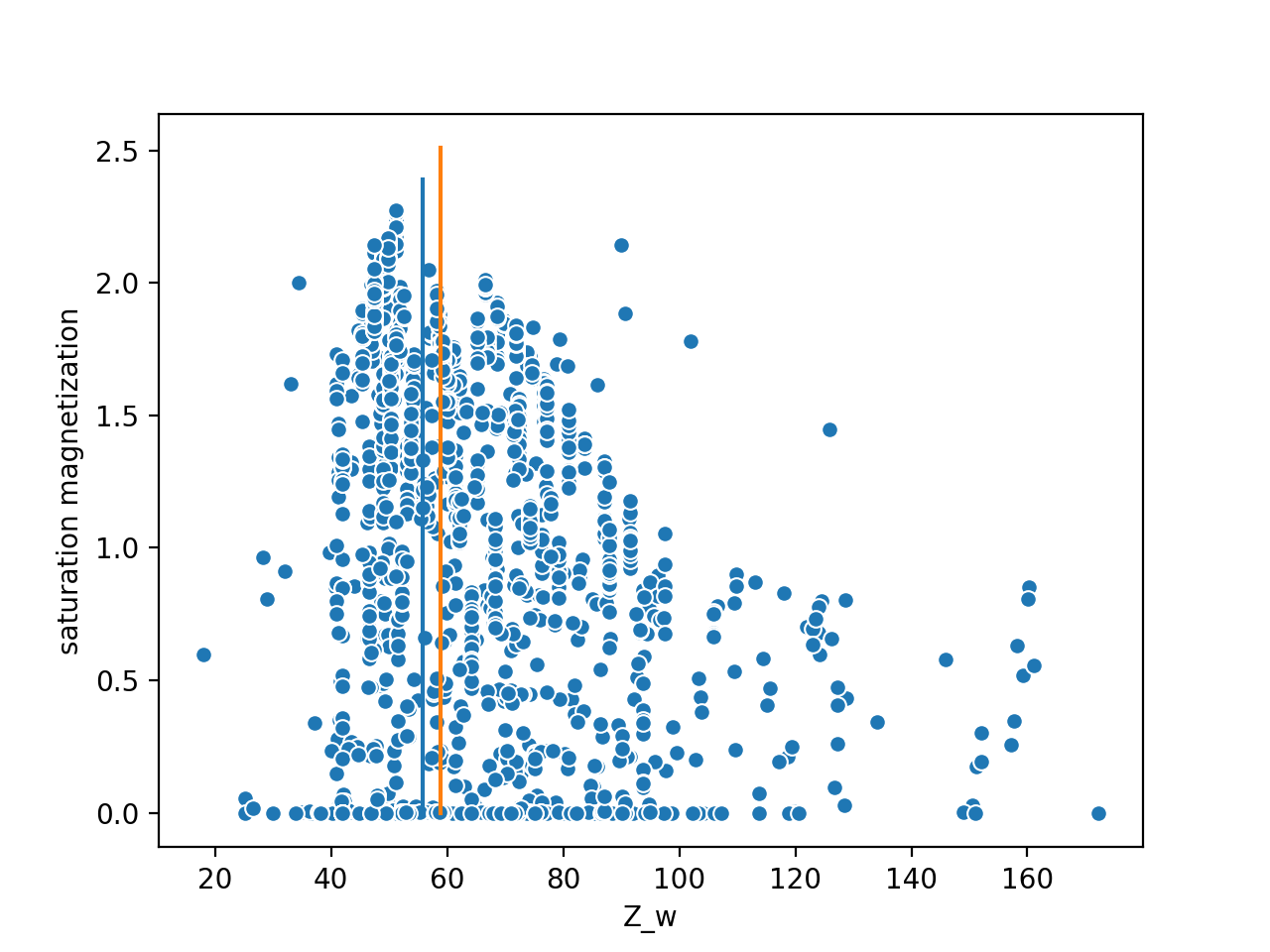
An example is the composition weighted atomic number:

e.g. for Fe8Co2N8 => (8/18)ZFe + (2/18)ZCo + (8/18)ZN

We can plot the distribution density when binned with this new feature. The orange and green lines indicate the atomic weight of Fe and Co respectively. The bimodal peak centred

around the transition metal ferromagnetics likely occurs due to combining these elements with elements with substantially higher and lower atomic weights.



Plotting instead the saturation magnetisation as a function of Z\_w, we find that the compounds with a weighted atomic mass of around 50 show the highest values. This suggests that of the two peaks in the previous figure, alloyed transition metal ferromagnets are more likely to maintain a higher saturation magnetisation when the alloy contains lighter elements.